AI Cyberdefense

As artificial intelligence (AI) progresses and embeds itself into every facet of digital life, its applications in cyberdefense become increasingly critical. This repository serves as a comprehensive compilation of resources pertaining to AI-based cyberdefense. It curates a wide spectrum of materials, ranging from books and articles on general cybersecurity and AI cybersafety, to the application of machine learning techniques in cybersecurity. Additionally, it provides an overview of prevalent cyber attacks and potential defenses, highlights relevant conferences and events, and introduces products and initiatives from leading AI organizations committed to strengthening cyber defense. By equipping readers with this knowledge, we aim to empower individuals, organizations, and nation-states to leverage AI technologies in fortifying their cyber infrastructure and effectively combat the rising tide of cyber threats. This work underscores the urgent need for informed and proactive engagement in this rapidly evolving landscape of AI and cyber defense.

Resources

General cybersecurity

- [list] Goodreads list of books (https://www.goodreads.com/review/list/72754976?shelf=cybersecurity&sort=avg_rating#)
    - [book] Silence on the Wire (https://cloudflare-ipfs.com/ipfs/bafykbzaced7qlzap77wrfegoup2djxbeafqeeasp47sgx563hmxhdnn4usfzg?filename=Michal%20Zalewski%20-%20Silence%20on%20the%20Wire_%20A%20Field%20Guide%20to%20Passive%20Rec No%20Starch%20Press%20%282005%29.pdf): It is supposedly one of the better introductions to "how the internet works"
- [post] Overview in Danish (https://www.bdo.dk/da-dk/services/advisory/cybersikkerhed?utm_source=bing&utm_medium=cpc&utm_campaign=Service%20-%20Cybersikkerhed%20(Dansk)&utm_term=Cyber%20Security&utm_content=Cyber%20Security)

- [report] MIT Cybersecurity review of 20 countries (https://www.technologyreview.com/2022/11/15/1063189/the-cyber-defense-index-2022-23/)
- [website] Live cyber threat map (https://threatmap.checkpoint.com/)
- [whitepaper] Checkpoint's guide for adopting a threat prevention approach to cybersecurity (https://pages.checkpoint.com/preventing-unknown-zero-day-attacks-whitepaper.html)
- [report] IBM's estimates for costs of data breaches (https://www.ibm.com/reports/data-breach)
- [article] Building a vulnerability benchmark (https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-ifs.2018.5647)
- [guidelines] NIST Framework for Improving Critical Infrastructure Cybersecurity (https://www.baltimorecityschools.org/sites/default/files/inline-files/NIST.CSWP_.04162018.pdf)
- [article] Social cybersecurity: an emerging science (https://link.springer.com/article/10.1007/s10588-020-09322-9)
- [textbook] Cybersecurity for Industry 4.0 (https://link.springer.com/book/10.1007/978-3-319-50660-9)

AI cybersafety

- [article] Review: machine learning techniques applied to cybersecurity (https://link.springer.com/article/10.1007/s13042-018-00906-1)
- [article] Cybersecurity data science: an overview from machine learning perspective (https://link.springer.com/article/10.1186/s40537-020-00318-5)
- [article] Machine learning approaches to IoT security: A systematic literature review (https://www.sciencedirect.com/science/article/pii/S2542660521000093)
- [sequence] AI infosec: first strikes, zero-day markets, hardware supply chains, adoption barriers (https://www.lesswrong.com/posts/kvk2ZorXui4YB4zvc/ai-infosec-first-strikes-zero-day-markets-hardware-supply)
- [post] AI Safety in a World of Vulnerable Machine Learning Systems (https://www.lesswrong.com/posts/ncsxcf8CkDveXBCrA/ai-safety-in-a-world-of-vulnerable-machine-learning-systems-1)

Other lists

- BlueTeam-Tools (https://github.com/A-poc/BlueTeam-Tools): This github repository contains a collection of 65+ tools and resources that can be useful for blue teaming activities.
- LLM-based cybersecurity tools (https://github.com/tenable/awesome-llm-cybersecurity-tools)
- RedTeam-Tools (https://github.com/A-poc/RedTeam-Tools): This github repository contains a collection of 130+ tools and resources that can be useful for red teaming activities.
- awesome-security (https://github.com/sbilly/awesome-security)
- Reseach-AI-CyberSecurity (https://github.com/AIDXNZ/Research-Ai-Cybersec): A collection of resources to start off researching AI in CyberSecurity
- Awesome Cyber Security (https://github.com/fabionoth/awesome-cyber-security): A collection of awesome software, libraries, documents, books, resources and cool stuff about security.
- Awesome AI for cybersecurity (https://github.com/Billy1900/Awesome-AI-for-cybersecurity): Awesome list of AI for cybersecurity including network (network traffic analysis and intrusion detection), endpoint (anti-malware), application (WAF or database firewalls), user (UBA), process behavior (anti-fraud).
- Awesome Machine Learning for Cyber Security (https://github.com/jivoi/awesome-ml-for-cybersecurity): A curated list of amazingly awesome tools and resources related to the use of machine learning for cyber security.
- Awesome AI Security (https://github.com/DeepSpaceHarbor/Awesome-AI-Security): A curated list of AI security resources inspired by awesome-adversarial-machine-learning & awesome-ml-for-cybersecurity.

Experimental resources

We're interested in running experiments on how we can make cybersecurity safer or increase the reliability and defense of LLM systems.

Datasets

- A really good overview of real-world networking datasets and resources (https://gist.github.com/stefanbschneider/96602bb3c8b256b90058d59f337a0e59)
- Darknet dataset (https://www.unb.ca/cic/datasets/darknet2020.html) (Δ (https://www.kaggle.com/datasets/peterfriedrich1/cicdarknet2020-internet-traffic))

Packages

- [python] FlowLabeler (https://github.com/jsrojas/FlowLabeler): Processing IP packets
- [python] Malware environment for OpenAI Gym (https://github.com/endgameinc/gym-malware): Create an AI that learns through reinforcement learning which functionality-preserving transformations to make on a malware sample to break through / bypass machine learning static-analysis malware detection.

References

- EvadeRL: Evading PDF Malware Classifiers with Deep Reinforcement Learning (https://www.hindawi.com/journals/scn/2022/7218800/)

Project ideas

- Malware detection: AI has the potential to provide much more accurate and faster detection of malicious activity than traditional signature-based detection. To design this kind of system, you would need to first create a data set of network traffic. This data set would need to include both malicious and benign traffic so that the AI could learn to distinguish between the two. | Network traffic dataset
- LLM Phishing Detection: Train an LLM to generate phishing emails and use it as a benchmark to train and test anti-phishing systems (dataset 1 (https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning), dataset 2 (https://github.com/GregaVrbancic/Phishing-Dataset)).
- Input Sanitization Check: Test various unsanitized inputs to an LLM and observe if it can be exploited to perform unintended operations, such as SQL injection or Cross-Site Scripting (XSS).
- Malicious Code Generation Prevention: Test different safety mitigations in preventing an LLM from generating harmful code snippets, even when specifically requested. This can involve testing various prompts and fine-tuning strategies.
- Safety Layers Benchmarking: Evaluate the effectiveness of various safety layers (rate limiters, use-case specific guidelines) in protecting the LLM from misuse.
- LLM Chatbot Resilience: Evaluate how well an LLM chatbot can withstand attempted attacks or malicious uses by simulating an adversarial user trying to trick the system into generating harmful content.
- Evaluating LLMs for Intrusion Detection: Test LLMs' capability to detect intrusion attempts in network traffic data, compared to traditional IDS systems.
- Exploit Generation Prevention: Evaluate the ability of LLMs to generate known software exploits when given a description of a vulnerability. The aim is to prevent the model from generating such exploits.
- Content Filtering Effectiveness: Evaluate the effectiveness of content filtering mechanisms in LLMs

in blocking the generation of malicious content.
- LLM Robustness to Adversarial Attacks: Test the robustness of an LLM to adversarial attacks, where inputs are deliberately crafted to mislead the model or cause it to generate malicious outputs.
- Differential Privacy Implementation: Implement differential privacy techniques to protect sensitive information in LLM training data and evaluate how this affects the model's ability to generate malicious content.

Conferences and events

Name	When?	Description	Location
44CON (https://44con.com/2023/03/20/44con-2023-early-bird-tickets/)	13-15 Sep 2023		London
CCCamp (https://events.ccc.de/camp/2023/infos/)	15-19 Aug 2023	A hacker camp	Berlin
SEC-T (https://www.sec-t.org/)	12-15 Sep 2023	Conf. w/ talks & Q&As	Stockholm

Products

- Palantir AIP (https://www.palantir.com/platforms/aip/)
- CIS Benchmarks (https://www.cisecurity.org/cis-benchmarks)
- Intel Owl (https://github.com/intelowlproject/IntelOwl): Single API to get threat information about any file, IP, etc.

AI organization commitments

- Anthropic Trust (https://trust.anthropic.com/)
- OpenAI Trust (https://trust.openai.com/)

Notes

Overview of attacks and defenses

Attack	Defense	Defense description
Malware attacks: Malicious software	Antivirus, antimalware (AMW) software, firewalls	AMW: Signature-based (known malware) and behaviour-based detection (suspicious activity).
Phishing attacks	User education, email filtering, network traffic flagging	
Denial-of-service attacks	Network capacity, CDN, intrusion detection & prevention system (IDPS)	IDPS monitors network traffic and warns or blocks